基于改进 SimRank 的产品特征聚类研究 *

刘　臣，段　俊†

(上海理工大学 管理学院，上海 200093)

摘　要：针对在线用户评论中产品特征的提取和聚类问题进行了研究，提出一种改进的 SimRank 算法，将情感词-特征对放入二分网中，在二分网中使用改进后的 SimRank 算法计算特征词之间的相似度；再通过谱聚类算法对特征相似度进行聚类，提取网络产品的特征集合。以某电脑评论为例，从中提取情感词-特征对进行研究，实验结果显示，改进后的算法准确率更高。改进后的特征相似度检测方法可以作为检测特征相似度的有效方法。实验采用在线产品的评论语料，实验结果表明使用改进后的 SinRank 相似度对特征词进行聚类提取出特征更加准确。

关键词：SimRank 算法；特征聚类；二分网；特征相似度

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.01.0027

Product feature clustering based on improved SimRank

Liu Chen, Duan Jun†

(*Business School*, *University of Shanghai for Science & Technology*, *Shanghai* 200093, *China*)

Abstract: This paper studies the extraction and clustering of product features in online user reviews. It proposed an improved SimRank algorithm to put the affective word-feature pair into the binary network. And the improved SimRank algorithm is used to compute the similarity between the characteristic words. Then the spectral clustering algorithm is adopted to cluster the feature similarity. Extracts feature sets for network products. Taking a computer commentary as an example, the paper extracts affective word-feature pairs, and the experimental results show that the improved algorithm has higher accuracy. The improved feature similarity detection method can be used as an effective method for detecting feature similarity. The experimental results show that using the improved Sinrank similarity to extract the feature words is more accurate.

Key words: SimRank algorithm; feature clustering; binary network; feature similarit

0　引言

随着电子商务的迅速发展，消费者更多的选择网上购物。同时，消费者在购买产品后会发表产品评论。互联网上涌现出大量的产品评论信息，这些评论对在线产品的各个性能进行了描述。研究表明，消费者会根据产品评论做出是否购买的决策。所以，产品评论对消费者和商家都至关重要。但是，消费者阅读在线评论存在很多阻碍。例如：a)评论信息过载,由于某些产品评论量过大，消费者无法一一阅读来获取自己所需信息，而大部分消费者往往只需要根据自己的需求了解产品的部分特征；b)用户评论中存在产品特征表达多样性问题，如“外观”，“外表”，“外形”和“颜值”等描述同一类特征，消费者很难通过读完所有评论了解到该产品的全部特征。商家也无法客观的提供产品的全部特征，尤其是由消费者进行产品体验后的特征；c)几乎每个消费者对产品特征的关注点都不同，消费者很难从大量的评论中快速找到自己所需的特征等。基于众多问题，目前的研究热点之一即为如何将有价值的信息从海量的评论中挖掘出来。产品特征的有效提取[1]是解决上述诸多问题的关键方法。通过对网上产品评论信息的挖掘与分析[1]，用户能清晰的认识产品的特征，从而做出购买决策，商家也能掌握消费者对产品的评价，从而对产品的某些方面进行改进。

目前，已有一些学者在产品评论的特征提取方面做出了研究。Hu 和 Liu[2]最开始使用关联规则实现对产品特征的提取。Kamal[3]等人设计出的相关规则对评论中产品“特征-情感对”进行抽取则是通过对评论的语义和语言学分析。在商品特征自动提取方面，Hu 等人[4]利用词性标注在评论中提取名词或名词性短语从而生成事务集，根据 Apriori 算法提取候选特征集，最后使用邻近规则和独立支持度规则过滤商品特征。李实等人[5]参考 Hu 等人的研究方法，结合中文评论的语言特点，对 Hu 等人的方法进行改进，将改进后的方法应用于提取中文评论中的产品特征。刘鸿宇等人[6]利用产品特征词与评价词之间的依存句法关系，同时结合频率、点互信息、名词剪枝三种过滤技术，

收稿日期：2018-01-17；**修回日期：**2018-03-07　**基金项目：**国家自然科学基金资助项目（71401107，71774111）

作者简介：刘臣（1981-），男，河南南阳人，副教授，博士，主要研究方向为互联网用户行为、Web 数据挖掘；段俊（1994-），女（通信作者），安徽马鞍山人，硕士研究生，主要研究方向为复杂网络、情感分析（18756217189@163.com）.

从在线评论中提取产品特征。Jin 等人[7]则直接从利弊型评论中提取出名词和名词短语，去除低频词及停止词后，利用 WordNet 词典进行同义词聚类，聚类后的集合作为产品特征列表。张珠等人[8]衡量商品特征之间关联程度是运用了语素和评价词，用 K-means 方法实现商品特征归类，但 K-means 聚类方法具有一些缺陷，如数据离群点、输入次序等会影响聚类质量。

因此，本文从复杂网络的角度对产品特征进行聚类分析，将提取出的特征词、情感词分别看作二分网络的顶点，情感词-特征对之间的关系看作二分网络的边。在 SimRank 算法的基础上考虑二分网络边的权重，分别使用原相似度计算方法和改进后的方法对样本数据进行相似度检测，并对比准确率验证改进后方法的有效性。最后利用谱聚类分别对得到的相似性矩阵及加权的相似度矩阵进行聚类。实验表明改进的 SimRank 算法计算特征词之间的相似性更加准确，聚类效果更好。

1　实验方法

1.1　数据抓取及分词标注

本文使用 Python 编写爬虫程序，在京东网上抓取产品评论[9]，抓取的评论保存到数据库中。从京东商城上抓取了某电脑的全部评论，共 11 543 条。

首先对爬取到的评论进行预处理[10]，使用哈工大社会计算与信息检索研究中心研发的“语言技术平台（LTP）”。LTP 中的中文自然语言处理[11]云服务高效精准。LTP 进行分析的前两步是将评论语句分词及词性标注，基于此，我们使用 Python 对存放在数据库中某电脑全部评论进行分词和标注。分词标注后的词性有名词 n，形容词 a，动词 v，副词 d，连词 c，标点符号 wp 等。然后通过分析分词标注后的各成分之间的依存语法关系揭示评论的句法结构。依存句法分析通过识别评论语句中的“主谓宾”、“定状补”这些语法成分来分析它们之间的关系。主要语法关系有主谓关系（SBV），动宾关系（VOB），定中关系（ATT），状中结构（ADV），并列关系（COO）等。如图 1 所示。

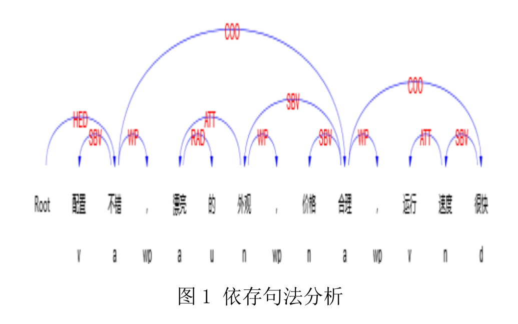


图 1　依存句法分析

1.2　提取情感词-特征对

将京东的产品评论爬取出来，根据 LTP 中的依存句法分析，找出特征词和情感词之间的依赖关系。通过大量的训练发现情感词和特征词之间主要语法关系主要有：多数的主谓关系（SBV），少数的动宾关系（VOB）、定中关系（ATT）以及并列关系（COO）。

因此，基于大量训练得出的特征词和情感词之间的关系，将预处理后的评论进行情感词-特征对提取[12]。提取出的情感词-特征对作为候选情感词-特征对。部分如表 1 所示。

表 1 依赖关系提取情感词-特征对

特征词	情感词	依赖关系
配置	不错	SBV
外观	漂亮	ATT
价格	合理	SBV
速度	很快	SBV
…	…	…

1.3　特征词相似度计算

本文将特征词与情感词之间的联系抽象为一个二分网络。由不同性质的两类节点以及这两类节点之间的连边组成二分网络，其中同类节点之间没有连线。在这个网络中，将情感词看作一类节点，特征词看做另一类节点，而用依存句法分析出的情感词与特征词之间的语法关系看成情感词节点与特征词节点之间的连边。在情感词-特征对的提取过程中，情感词和特征词之间的关系就呈现出这样的网络特征。如图 2 所示。

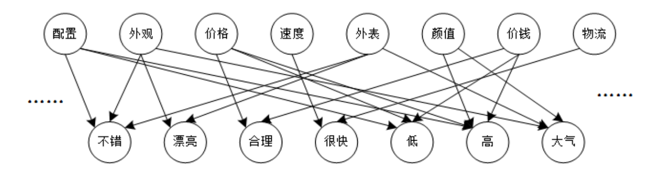


图 2　二分网络图

基于情感词-特征对二分网络，本文使用 SimRank 相似度算法计算特征词之间的相似度[13]。SimRank 相似度算法的核心思想为：如果两个对象和被其相似的对象所引用（即它们有相似的入邻边结构），那么这两个对象也相似。近年来已在信息检索领域引起广泛关注，成功应用于网页排名、网络图聚类、协同过滤、孤立点检测、近似查询处理等。SimRank 基本公式：

$$S(a,b)=\begin{cases}1, a=b\\ \frac{c}{|I(a)||I(b)|}\sum_{i}^{|I(a)|}\sum_{j}^{|I(b)|}s\left(I_i(a),I_j(b)\right), a\neq b|I(a)|,|I(b)|\neq\emptyset\\ 0, otherwise\end{cases}$$

（1）

其中：$S(a,b)$ 指节点a，节点b之间的相似度，取值范围是[0,1]；$I(a)$为所有指向结点a的节点集合，即入邻点集合；$I_i(a)$ 为第i个指向a的节点，即a的第i个入邻点；$s\left(I_i(a),I_j(b)\right)$为相连的二部图另一个子集节点之间的相似度，即相连的情感词之间的相似度。参数c是阻尼系数，$0<c<1$，一般取$c=0.8$。所以，节点a与节点b之间的相似

度就等于所有同时指向节点a，b的节点相似度之间的平均值。

由于 SimRank 算法认为只要有边相连，即为相似。如图 3 所示，相似的特征词都和相同的情感词相连接，而非相似的特征词也会和同一情感词连接，但是没有考虑到两个节点共同相连的边越多，则意味着节点相似度越高。实验数据也证明直接使用 SimRank 算法计算的特征词之间的相似度无太大差别，所以仅通过 SimRank 求出的相似度，对这些数据进行聚类效果不好。所以，这里使用改进的 SimRank 算法，本文考虑加入权重的思想，即在算法中加入二分网边的权重。

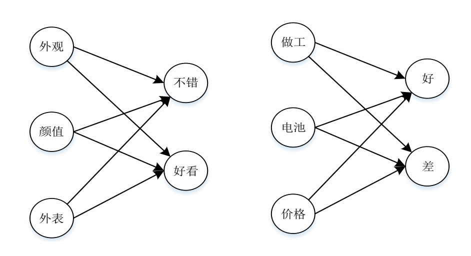


图 3 情感词-特征对二分网络图

1.4　情感词-特征对边权重计算

基于二分网的特征计算情感词-特征对边权重。首先统计出二分网边的频次f，即情感词-特征对出现的频次，进而计算每条边的两个端点出现的概率：

$$p_1 = \frac{f_1}{\sum_{i=1}^{n} f_i} \quad （2）$$

$$p_2 = \frac{f_2}{\sum_{j=1}^{n} f_j} \quad （3）$$

其中：f_i表示该边连接的特征词出现的频次，f_j表示该边连接的情感词出现的频次，p_1，p_2分别表示特征词和情感词出现的概率。

在此基础上计算每条边的特征词和情感词两端点所携带的信息熵：

$$I(p_1) = -p_1 \log_2 p_1 \quad （4）$$

$$I(p_2) = -p_2 \log_2 p_2 \quad （5）$$

为了计算的更加准确，我们用下式作为边的权重，即

$$w = (-p_1 \log_2 p_1 - p_2 \log_2 p_2) \log_2 f \quad （6）$$

1.5　加权特征词相似度计算

SimRank 相似度算法中，由于两节点共同相连的边越多意味着节点相似度越高，本文在二分

网中加入权重来计算特征词之间的相似度。即改进的 SinRank 相似度算法。加权的 SimRank 迭代算法：

$$R_0 = (a,b) = \begin{cases} 0, a \neq b \\ 1, a = b \end{cases} \quad （7）$$

$$R_{k+1}(a,b) = \begin{cases} \frac{c}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k\left(I_i(a), I_j(b)\right) w(I_i(a) \to a) w(I_i(b) \to b), a \neq b \\ 1, a = b \end{cases} \quad （8）$$

其中：$w(I_i(a) \to a)$表示与a相连的第i个入邻点之间边的权重。二分网络中加权的 SimRank 公式：

$$S(A,B) = \frac{C_1}{|O(A)||O(B)|} \sum_{i=1}^{|O(A)|} \sum_{j=1}^{|O(B)|} S\left(O_i(A), O_j(B)\right) w(O_i(A) \to A)\, w\left(O_j(B) \to B\right), A \neq B \quad （9）$$

其中：w为情感词-特征对边的权重；节点A和B之间的相似度即为A，B的所有出度间加权相似度的平均值。同理，在情感词-特征对的二分网中A，B代表特征词，特征词 A，B之间的相似度等于指向A，B的所有情感词之间加权相似度的平均值。

2　数据分析

本文在 Python 中编写爬虫程序，从京东商城中抓取某电脑的全部在线评论。LTP 语言云对每条评论分词标注，提取情感词-特征对，首先根据 SimRank 相似度计算各特征词之间的相似度。部分结果如表 2 所示。

表 2 未加权的特征词相似度

SimRank	卡	电脑	屏幕	本本	外观	外形	…	笔记本	快递	物流	显卡
卡	–	0.003	0.0046	0.005	0.016	0.04	…	0.04	0.0133	0.0067	0.005
电脑	0.003	–	0.0027	0.003	0.0036	0.003	…	0.0059	0.004	0.0049	0.003
屏幕	0.0046	0.0027	–	0.0015	0.0049	0.002	…	0.0062	0.0041	0.0031	0.0038
本本	0.005	0.003	0.0015	–	0.012	0.0267	…	0.002	0.0067	0.0033	0.005
外观	0.016	0.0036	0.0049	0.012	–	0.0091	…	0.032	0.0107	0.0107	0.008
外形	0.04	0.003	0.002	0.0267	0.0091	–	…	0.023	0.008	0.008	0.003
…	…	…	…	…	…	…	…	…	…	…	…
笔记本	0.04	0.0059	0.0062	0.002	0.032	0.023	…	–	0.0533	0.0267	0.02
快递	0.0133	0.004	0.0041	0.0067	0.0107	0.008	…	0.0533	–	0.0267	0.0133
物流	0.0067	0.0049	0.0031	0.0033	0.0107	0.008	…	0.0267	0.0267	–	0.002
显卡	0.005	0.003	0.0038	0.005	0.008	0.003	…	0.02	0.0133	0.002	–

从实验后的相似度数据可以看出，由于表中数据的鉴别指数$D = P_h - P_d$，即最大值与最小的差为 0.0513。并且衡量数据离散程度的方差为 0.00014，都非常小，所以未加权的特征词之间的相似度区分度不高。同类特征词之间的相似度与不同类特征词之间的相似度差别不大。

对表 2 中的数据加权后得到表 3 中的数据，鉴别指数为

0.8871，方差为 0.06718。相比较表 2 中的数据，表 3 中的数据有了很大的区分度。如表 3 中加黑部分的相似度值，同类特征词的相似度明显高于不同类的特征词，实验数据便于做后续的聚类分析。

表 3 加权的特征词相似度

WSimRank	卡	电脑	屏幕	本本	外观	外形	…	笔记本	快递	物流	显卡
卡	–	0.1012	0.0582	0.2754	0.1937	0.0110	…	0.0173	0.3337	0.0115	**0.8046**
电脑	0.1012	–	0.1363	**0.7059**	0.0655	0.0353	…	**0.7934**	0.3234	0.0324	0.0555
屏幕	0.0582	0.1363	–	0.0071	0.1170	0.0544	…	0.4755	0.0112	0.0261	0.0390
本本	0.2754	**0.7059**	0.0071	–	0.0675	0.0204	…	**0.6784**	0.3131	0.0279	0.0085
外观	0.1937	0.0655	0.1170	0.0675	–	**0.8250**	…	0.5080	0.2761	0.0617	0.0241
外形	0.0110	0.0353	0.0544	0.0204	**0.8250**	–	…	0.2335	0.0878	0.0322	0.0243
…	…	…	…	…	…	…		…	…	…	…
笔记本	0.0173	**0.7934**	0.4755	**0.6748**	0.5080	0.2335	…	–	0.2845	0.0077	0.1847
快递	0.3337	0.3234	0.0112	0.3131	0.2761	0.0878	…	0.2845	–	**0.8942**	0.0402
物流	0.0115	0.0324	0.0261	0.0279	0.0617	0.0322	…	0.0077	**0.8942**	–	0.0335
显卡	**0.8046**	0.0555	0.0390	0.0085	0.0241	0.0243	…	0.1847	0.0402	0.0335	–

3 基于相似度聚类

本文给出了情感词-特征对边的相似度以后，特征词种类的划分就可以转化为聚类问题，基于相似度矩阵的聚类算法进行划分。在复杂网络的网络簇结构中存在着同族节点之间连接密集，不同节点之间连接稀疏的特征。根据这样的特征对网络节点进行聚类，使得同类节点之间连接密集，不同类节点之间连接稀疏。

这里使用谱聚类算法[14]对网络节点中的特征词进行聚类[15]。这里谱聚类是根据特征词之间的相似度，将它们分成不同组。根据谱聚类的思想，将特征词看作顶点V，特征词之间的相似度看作带权的边E，得到一个基于相似度的无向加权图$G(V,E)$，两点之间有边相连权重大于零，无边相连权重等于零，从而将聚类问题转为图分割问题。如图 4 所示。谱聚类能找到一种最优的图分割的方法，该方法使分割后的子图内部有最大的相似度，子图之间有最小的相似度。也就是不同组之间相连的边权重尽可能低，即组间相似度尽可能低，各个组内部相连的边权重尽可能高，即组内相似度尽可能高。

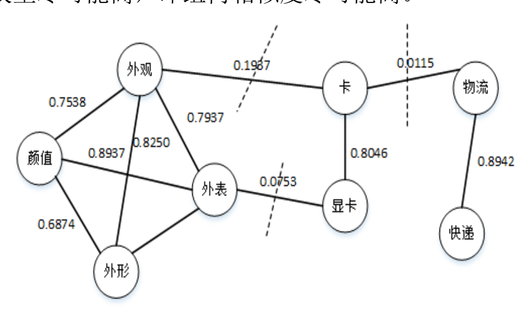


图 4 聚类原理图

实现方法的总体框架可以归纳为三个主要步骤：

a) 构建表示对象集的无向加权图，即相似度矩阵W，同时指定需要聚类的簇数k。根据前面加权的 SimRank 算法得到特征词之间的相似度矩阵。所以该聚类中无需再计算相似度矩阵，直接进行后半部分的步骤即可。且该相似度矩阵W具有如下性质：矩阵$N \cdot N$，为特征词总数；矩阵对角线的值为 0；矩阵为对称矩阵，即相似度是无向的。此时，谱聚类的任务就是根据这个相似度矩阵W，将这N个特征词分成不同的组，小组内部的特征词彼此都很像，小组之间则不像。通过这个相似度矩阵得到对应的图分割方案。

b) 计算对角矩阵$D(N \cdot N)$：

$$D_{i,i} = \sum_j^N w_{i,j} \qquad (10)$$

对角矩阵D对角线上的值为矩阵W中对应行或列的和，指的是度矩阵，即和点i相连的所有边的权重之和，$w_{i,j}$指得是第i行对应第j点的度数。为了有利于后面子图内部的极大化，利用相似度矩阵构造拉普拉斯矩阵L：$L = D - L$。再将拉普拉斯矩阵标准化，即归一化L矩阵：

$$L = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \qquad (11)$$

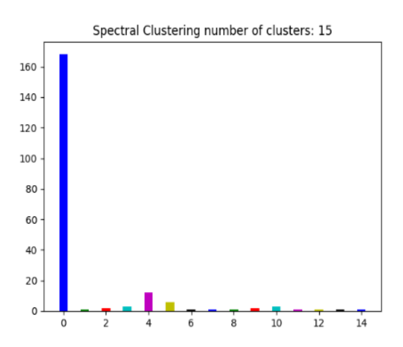
为了将L降维，计算L矩阵的K个最小特征值以及对应的特征向量，并将K个特征向量按列排放，形成$N \cdot K$的特征矩阵。

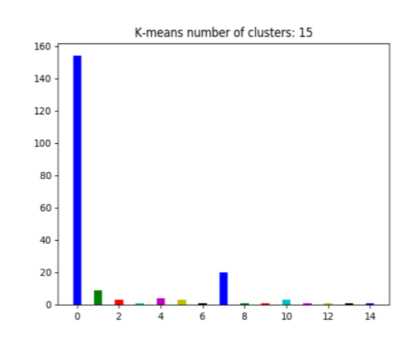
c) 对这里的特征矩阵进行 K-means 聚类，得到的N维向量分别对应相似度矩阵W中每一行所代表的特征词所属类别，也即最终聚类结果。

4 实验结果与分析

如图 5 所示，横轴表示聚类的个数，纵轴表示每个聚类包

含的个数，即每个聚类包含的节点数量。第一张图为聚类簇数 k 取 15 时未加权 SimRank 相似度矩阵聚类出的结果，第二张图是用传统的 K-means 聚类后的结果，最后一张图是本文中改进的加权相似度矩阵聚类出的结果。由图可以看出，未加权的相似度矩阵几乎无法进行聚类。相比之下，最后一张图的聚类效果要更加明显。且和传统的 K-means 聚类相比较聚类效果也更好。





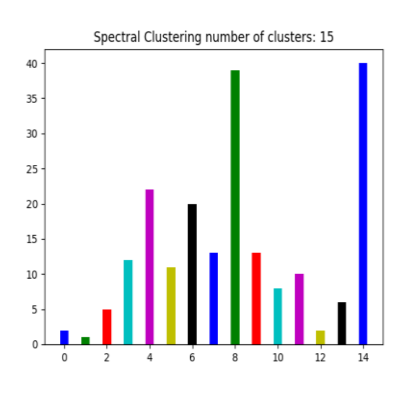


图 5 相似度聚类图

这里从每一行聚类分布的结果中得到聚类分布情况，电脑特征聚类部分结果如表 4 所示，第一列表示该簇的簇中心，第二列是该簇的簇内成员。从表中可看出，谱聚类效果较好且不受离群点的影响，如第二簇中的“做工”可单独作为一个簇。实验结果显示，本文所用方法准确率较高，更加适合特征词聚类。

表 4 特征词列表样例

簇中心	簇内成员
卡	显卡
做工	
笔记本	本本，本子，电脑 ……
外形	颜值，外观，外型，外壳 ……
音质	音效，音，音色，音响效果，声音 ……
物流	快递，发货，送货 ……
价格	价位，秒价，价钱，特价 ……
字	字体，文字，字号 ……
手感	质感，触感，塑料感，印刷感，体验感，动感……
性能	流畅性，扩展性，便携性，兼容性 ……
配置	配件，适配器，配器 ……
内存	容量，大小，内存条 ……
音响	音箱
插口	口，接口 ……
机器	机，一款机，机子，机身，真机，新机子，主机 ……
⋮	⋮

5　结束语

大数据时代海量的文本信息需要进行有效的处理，而特征词挖掘是文本信息处理的基础。针对产品评论挖掘产品特征进行了研究。本文首先通过依存句法实验训练出特征词与情感词之间的语法关系，并通过语法关系提取出某在线产品评论的全部情感词-特征对；然后将情感词-特征对放入二分网中，基于二分网使用 SimRank 算法计算特征相似度；为实验结果更加准确，提出基于改进的 SimRank 相似度出发，即加权特征相似度。基于商品特征相似度的聚类算法有效地对相近的商品特征进行聚类，成功降低了商品特征维度，为评论挖掘的进一步研究奠定基础。实验表明，本文所提出的改进后的特征词相似度计算方法能有效提高聚类准确度。同时也存在一些不足，如通过 LTP 提取出的情感词-特征对并非完全准确，这给后续的相关研究工作带来不便。针对存在的不足之处，将做进一步研究。

参考文献：

[1] Liu Bing. Sentiment analysis and opinion mining [M]. Morgan: Claypool Publishers, 2012: 67-77.

[2] Hu Minqing, Liu Bing. Mining and summarizing customer reviews [C]// Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 168-177.

[3] Kamal A, Abulaish M, Anwar T. Mining feature-opinion pairs and their reliability scores from web opinion sources [C]// Proc of the 2nd International Conference on Web Intelligence, Mining and Semantics. 2012:

1-7.

[4] Hu Minqing, Liu Bing. Mining opinion features in customer reviews [C]// Proc of the 19th International Conference on Artifical Intelligence. San Jose: AAAI Press 2004: 755-760.

[5] 李实，叶强，李一军，等. 挖掘中文网络客户评论的产品特征及情感倾向 [J]. 计算机应用研究, 2010, 27 (8): 3016-3019. (Li Shi, Ye Qiang, Li Yijun, *et al.* Mining Product Features and Sentiment Orientation from Chinese Customer Reviews [J]. Application Research of Computers, 2010, 27 (8): 3016-3019.)

[6] 刘鸿宇，赵妍妍，秦兵，等. 评价对象抽取及其倾向性分析 [J]. 中文信息学报, 2010, 24 (1): 84-88.

[7] Jin J, Liu Y, Ji P, *et al.* Understanding big consumer opinion data for market-driven product design [J]. International Journal of Production Research, 2016, 54 (10): 3019-3041.

[8] 张珠，贾文杰，夏迎炬，等. 商品属性归类技术研究 [C] //第六届全国信息检索学术会议论文集. 2010.

[9] Shi B, Chang K. Mining Chinese Reviews [C]// Proc of the 6th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2006: 585-589.

[10] Xu H, Zhang F, Wang W. Implicit feature identification in Chinese reviews using explicit topic mining model [J]. Knowledge-based systems, 2015, 76 (5): 166-175.

[11] Yi J, Nasukawa, Bunescu R, etal. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques [C]// Proc of the 3rd IEEE International Conference on Data Mining. Washington D C: IEEE Computer Society, 2003: 427.

[12] Zhang S, Jia W, Xia Y, *et al.* Product features extraction and categorization in Chinese reviews [C]// Proc of the 6th IEEE International Conference on Advanced Information Management and Service. 2010: 324-329.

[13] Glen J, Widom J. SimRank: a measure of structural-context similarity [C]// Proc of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 538-543

[14] Li S, Hao J. Spectral clustering-based semi-supervised sentiment classification [C]// Advanced Data Mining and Applications. Berlin: Springer, 2012: 271-283.

[15] Pang B, Lee L. Opinion mining and sentiment analysis [J]. Foundations and Trends in Information Retrieval, 2008, 2 (1//2): 1-135.